Marine Pollution Bulletin 64 (2012) 2871–2876

Contents lists available at SciVerse ScienceDirect

Marine Pollution Bulletin

journal homepage: www.elsevier.com/locate/marpolbul

Baseline

Field studies using fish biomarkers – How many fish are enough?

M.M. Gagnon [a,*], P.V. Hodson [b]

[a] Department of Environment and Agriculture, Curtin University, P.O. Box U1987, Perth, Western Australia 6845, Australia
[b] School of Environmental Studies, Queens University, Kingston, Ontario, Canada K7L 3N6

ARTICLE INFO

Keywords:
Fish biomarkers
Sample size
Statistical significance
Variability
EROD
Biliary metabolites
LSI
GSI
CF

ABSTRACT

Significant efforts are invested in field studies using fish, and it is important to optimize the number of organisms collected to evaluate the possible impacts of contamination. This paper provides ecotoxicologists with the approximate numbers of fish needed to identify statistically significant differences among samples using physiological indices and biochemical markers of fish health. The numbers of fish to collect are reported for ethoxyresorufin-*o*-deethylase (EROD); ethoxycoumarin-*o*-deethylase (ECOD), serum sorbitol dehydrogenase (sSDH), stress proteins, gonadosomatic index, liver somatic index, condition factor, and biliary metabolites of polycyclic aromatic hydrocarbons. The number of fish to collect was as few as four for ECOD activity (with a power of 80%), but as high as 106 for CF (with a power of 95%). Achieving statistical significance between sites does not help in the interpretation of the biological significance of a parameter, but well-planned field samplings will maximize the chances of correctly identifying areas of concern.

1. Introduction

Most field programs that monitor chemical effects on fish compare the characteristics of fish captured at reference sites to those of fish collected at impacted sites. Sampling sites are usually selected to maximize the probability of detecting statistical differences between reference and impacted locations. Because field sampling requires significant financial and logistic efforts, it is important to optimize the number of organisms collected to evaluate the possible impacts of contamination with the lowest effort and cost. The appropriate number of specimens to collect should be determined for each sampling program, keeping in mind that field collection is often by far the most expensive part of a monitoring program.

Fish have proven useful as sentinel organisms which display measurable biological responses (biomarkers) that vary in proportion to the extent of exposure to contaminants. For example, the induction of ethoxyresorufin-*o*-deethylase (EROD) activity is one of the most popular biomarkers of exposure to aquatic contaminants such as polycyclic aromatic hydrocarbons (PAH). Consequently, the number of fish needed to establish significant inter-site differences in EROD activity has been the subject of several publications (e.g., Flammarion and Garric, 1997, 1999; Beliaeff and Burgeot, 1997; Oris and Roberts, 2007). EROD activity, is not the only response assessed to evaluate the health status of fish populations. However, each biomarker may demonstrate a unique variability and require a different number of specimens to establish inter-site differences. Information on the required number of samples to establish a significant difference for biomarkers other than EROD activity is practically non-existent in the literature.

The first intent of the present study is to provide ecotoxicologists with an approximation of the sample sizes required to detect a biologically relevant and statistically significant difference between sites for several biomarkers frequently measured in field-collected fish. It is well understood that sample size is a function of the degree of inter-species and inter-site differences, and the variability of the measurement. Therefore, the magnitudes of the inter-site differences within one species have been estimated from the literature to represent, or to be associated with, biologically relevant effects for individual fish or fish populations. We examined sources of variability in measured biomarkers, with a focus on EROD activity. The second intent of this paper is to provide a clear procedure for calculating required sample sizes for biologists who use statistics as a tool rather than as a mainstream science.

2. Materials and methods

To estimate minimum sample sizes, existing data are required to provide an initial estimate of variability. Where no pre-existing data are available, data extracted from similar field studies might be used, the variability of a biomarker and the magnitude of change between reference and impacted sites might be estimated from laboratory studies, or a small-scale preliminary field collection could

* Corresponding author. Tel.: +61 8 9266 3723; fax: +61 8 9266 2495.
E-mail addresses: m.gagnon@curtin.edu.au (M.M. Gagnon), peter.hodson@queensu.ca (P.V. Hodson).

0025-326X
http://dx.doi.org/10.1016/j.marpolbul.2012.08.016

be conducted. For this paper, we drew data from an existing survey of a contaminated estuary (Webb et al., 2005a,b).

Urban contamination has caused the health of the Swan River Estuary, Western Australia, to deteriorate significantly over time. To evaluate the health of the fish populations living in this system, a large field study was undertaken, in which black bream (*Acanthopagrus butcheri*) were sampled at several sites over time and tissues collected for biomarker analyses (see Webb et al., 2005a,b for methods and results). This study provided a large data set for a suite of biomarker results from 20 adult fish per site from four sites. These fish were collected during the inter-spawning period when they were not reproducing. Only the first 20 fish sampled within one season and with a complete set of biomarker data were included in this data set, for a total of 80 fish from the four locations in the estuary.

No true reference site exists in the Swan River Estuary, as the entire estuary has been impacted by human activities. There are still, however, some areas where impacts of non-nutrient contaminants are minimal, which we have defined as reference areas. Consequently, the four sampling sites included a reference, a highly impacted, and two intermediate-effect sites.

Biomarkers measured on the black bream included: EROD activity, ethoxycoumarin-*o*-deethylase (ECOD) activity, serum sorbitol dehydrogenase (sSDH) activity, naphthalene-, pyrene-, and B(a)P-type biliary metabolites, stress proteins (HSP70), liver somatic index (LSI = (liver weight/carcass weight) $\times$ 100), gonado-somatic index (GSI = [gonad weight/carcass weight] $\times$ 100), and condition factor (CF = carcass weight/length^3). While EROD activity and biliary PAH metabolites in fish have been identified as some of the most valuable and reliable biomarkers for risk assessment (van der Oost et al., 2003; Jung et al., 2011), the selected suite of biomarkers must be relevant to the case study.

The selection of a minimum detectable difference requires a consideration of biological significance. Biologically significant inter-site differences might be characteristic of each biomarker and each species, as well as individual variability among fish collected at the same site and time. For each biomarker, a review of published studies established what magnitude of effect could be considered a biologically significant difference between reference and impacted fish. If the biomarker was considered solely an indicator of exposure rather than of effects, an association (not a cause and effect relationship) with an overall health effect was made (Table 1). Individual studies might discover different magnitudes and directions of biomarker responses according to the specific situation investigated.

Most biological field data require log-transformations to achieve normality and homogeneity of variances; consequently all biochemical measures presented here have been log-transformed based on preliminary tests of normality and homogeneity of variance. It is important to understand that in absolute terms, the difference sought between reference and impacted groups would be much greater for an induction than an inhibition. If we consider for example an enzymatic change with untransformed data, a 3-fold induction of activity represents a much larger absolute change than a 3-fold inhibition of activity. However, the proportional difference is identical. The required number of fish computed in the present exercise takes into account inhibition or induction of a parameter as all data were log-transformed prior to calculations.

Using the existing data from black bream (Table 2) (Webb et al., 2005a,b), the number of fish required to detect an inter-site difference at $\alpha = 0.05$ was calculated using the publicly available program G*Power 3.1.3 (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/). The following criteria were selected: '*F*-tests', 'ANOVA: fixed effects, omnibus, one way', and '*a priori* compute required sample size – given α, power and effect size'. Raw data were log-transformed to compute an ANOVA and obtain the necessary 'SD σ within group' (square root of error within groups), along with the average of each group to be compared, to determine the 'effect size *f*'. Calculations were performed for powers of 0.80 and 0.95, corresponding respectively to 80% and 95% chances of obtaining a significant difference among groups at $\alpha = 0.05$.

The minimum required number of fish was calculated for a minimum biologically relevant amplitude of change, according to published literature (Table 1). For a given biomarker, the logged values of the existing reference data were used to compute the reference site average, and the anti-log of this average was multiplied by the desired amplitude – then logged again as the impacted site average. For example, if the reference log(EROD) was 0.967 and the desired amplitude of change to detect was a 3-fold induction in EROD activity at the impacted sites, then the antilog of 0.967 was obtained by $10^{0.967} = 9.928 \times$ 3-fold induction = 29.80, log(29.80) = 1.444. This value of 1.444 was used as the log impacted site average, representing a 3-fold induction relative to the reference data.

Table 1
For individual biomarkers, justification of confidence intervals used in calculating number of fish required.

Parameter	Minimum change	Justification	References
EROD activity	3x	Higher activity associated with reduced general health index; high EROD activity co-occurring with low GSI and plasma estradiol	Adams et al., 1996; Mower et al., 2011
ECOD activity	3x	Moderately contaminated sites show up to 2.9x induction relative to uncontaminated sites.	Holdway et al., 1994
Serum SDH	2x	Elevated serum SDH activity related to liver damage and reduced liver metabolic activity; elevated serum SDH preceded increases in LSI and liver hyperplasia.	Holdway et al., 1994; Webb and Gagnon, 2007
Biliary metabolites (naphthalene, pyrene, BaP)	4x[a]	Increased biliary metabolites co-occur with altered condition factors in contaminated environments.	Escartin and Porte, 1999
HSP70	2x	A 1.5-fold increase in HSP70 level co-occurred with reduced plasma glucose, indicative of altered hepatic metabolic activity.	Vijayan et al., 1998
LSI	0.1x	Chronic exposure of fish to BKME effluent caused a 15% increase in LSI.	Huuskonen and Lindström-Seppa, 1995
GSI	0.1x	A reduction of 9.6% in gonad size has been associated with reduced growth and delayed age at first maturity in BKME[b]-exposed fish.	Gagnon et al., 1995
CF	0.1x	Increased (8–10%) adult CF associated with altered growth as measured by increased length at young ages.	Adams et al., 1996

[a] Although laboratory studies can demonstrate a 100-fold increase in PAH biliary metabolites following exposure to petroleum compounds, field studies often show much weaker differences between reference and contaminated sites (e.g. Gagnon and Holdway, 2002).
[b] BKME: bleached kraft mill effluent.

Table 2
Parameters observed at each site (untransformed values). The sample size was 20 fish per site.

						Coefficient of variation (cv, %)			
	Parameter	Site 1 (Ref)	Site 2	Site 3	Site 4	Site 1 (Ref)	Site 2	Site 3	Site 4
EROD activity (pmol/mg pr/min)	Arithmetic mean	14.6	25.6	42.2	22.3	86	65	54	127
	Std dev	12.6	16.8	22.6	28.3				
ECOD activity (pmol/mg pr/min)	Arithmetic mean	2.07	3.59	5.39	3.50	43	37	24	37
	Std dev	0.891	1.34	1.27	1.29				
sSDH activity (mIU/min)	Arithmetic mean	44.4	98.8	93.5	77.3	27	23	47	37
	Std dev	12.0	22.8	44.4	28.6				
Naphth. bile metabolites (mg/mg pr)	Arithmetic mean	30.2	136	285	65.0	22	19	27	38
	Std dev	6.63	26.3	77.6	24.4				
Pyrene bile metabolites (μg/mg pr)	Arithmetic mean	8.59	27.4	42.7	14.8	43	34	27	23
	Std dev	3.65	9.36	11.4	3.41				
B(a)P bile metabolites (μg/mg pr)	Arithmetic mean	27.7	38.8	112	35.6	38	50	27	36
	Std dev	10.5	19.5	30.0	12.9				
HSP70 (pixel/μg pr)	Arithmetic mean	357	752	722	553	50	49	40	45
	Std dev	179	367	292	248				
LSI	Arithmetic mean	1.24	1.54	1.67	1.54	29	12.9	23	23
	Std dev	0.354	0.198	0.388	0.353				
GSI	Arithmetic mean	1.54	1.30	1.23	1.57	28	33	24	33
	Std dev	0.429	0.431	0.298	0.518				
CF	Arithmetic mean	1.72	1.81	1.63	1.83	6.9	8.3	8.4	9.0
	Std dev	0.102	1.51	0.136	0.165				

Note: Aritmetic mean ± standard deviations are provided as information only. Data sets needed to be log-transformed for statistical analysis due to heteroscedasticity of frequency distributions. Units: EROD, ECOD: pmol/milligram protein/min; sSDH: milli-International Unit/min; naphthalene bile metabolites: mg/mg proteins; pyrene and BaP bile metabolites: μg/mg protein; HSP70 (heat shock proteins 70): pixels/μg proteins.

3. Results and discussion

For the Swan River estuary black bream, the minimum number of fish required to define a statistically significant difference for a pre-selected degree of change at $\alpha = 0.05$ ranged from <4 to >106 (Table 3). The calculated sample sizes varied according to the parameter considered, the variability of the parameter, and the size of the difference to be detected. The parameters requiring the fewest fish (4–16 fish per site) were EROD and ECOD activity, serum SDH, and biliary PAH metabolites. Analysis of HSP70, LSI, GSI and CF required considerably more fish per site (13–106). These numbers generally increased in direct proportion to requirements for decreasing amplitudes of the difference from reference values.

For EROD and ECOD activity, only 4–12 fish/site were needed to detect a 3-fold induction. Previous studies with other fish species gave similar results. Flammarion and Garric (1999) estimated that 13 fish/sex/season/site were required to detect a 2-fold induction of EROD activity at $\alpha = 0.05$ in chub (*Leuciscus cephalus*). Similarly, Beliaeff and Burgeot (1997) calculated for a variety of fish species that 10 fish were required to detect a 3-fold EROD activity induction at $\alpha = 0.10$. The required number of fish computed in the present investigation was comparable to numbers reported in the published literature for field studies, where EROD activity is, on average, investigated using $n = 7$ fish per site (and laboratory studies use on average five fish per treatment, Oris and Roberts, 2007). Some acute field exposures may cause large and significant difference with very few fish. For example, following an oil spill, a significant EROD induction in rockfish (*Sebastes schlegeli*) and in marbled flounder (*Pseudopleuronectes yokohamae*) was detected using only $n \geqslant 3$ fish per site (Jung et al., 2011).

The field sampling from which the black bream data set was extracted was conducted outside of the reproductive season for this species to avoid a gender bias in EROD activity. While EROD activity is unbiased by gender in this case, other parameters such as GSI and reproductive parameters in general could not be investigated properly using this data set because the fish were not sexually mature. While a 10% change in these parameters required that 43–106 fish be sampled, the field data suggest that only 13–36 fish per site would be sufficient, as inter-site differences in LSI and GSI often varied by more than 10%.

Four factors will influence the required number of samples (n) to collect. The first, the significance level α, is almost uniformly accepted at $\alpha = 0.05$, meaning that for 1 in 20 comparisons, there may be a false positive and incorrect conclusions about effects. Lowering α causes n to increase dramatically but it may be practical to collect a larger number of samples if the biomarker analyses are inexpensive, or if more fish are needed for other responses.

The second factor is the desired minimum detectable difference amongst sites, which will be specific to each location and to each biomarker. No obvious rulings exist for the magnitude of change that can be appropriate to specific situations (Hanson et al., 2010). For each biomarker, we estimated a biologically or environmentally relevant degree of change between reference and impacted fish (Table 1). Similarly, each monitoring program should identify a degree of change relevant to its own situation to calculate an optimal number of samples. While Table 1 lists the minimum change that could be associated with biologically relevant endpoints, other field studies have reported much higher changes in observed parameters. For example, populations of white sucker (*Catostomus commersoni*) exposed to bleached kraft mill effluents had GSI, LSI and CF deviations of 30% or more relative to reference fish (Mower et al., 2011).

The power of the test, 1-β, is a third factor influencing the number of samples to collect. The convention in environmental sciences is that power should be at least 0.80 (Fairweather, 1991), i.e., there should be an 80% chance of detecting a difference between sites. The power of a test can be determined easily from calculations using similar variables as the minimum sample size (G*Power 3 can calculate power using a different set of instructions). Obviously, collecting the minimum number of samples will give low power and increase the chances of committing a Type II error (false negative: concluding there is no impact when in fact there was one). In a multi-sample analysis of variance, the power increases rapidly with the number of samples used. Consequently, if there is an opportunity to collect a few more fish at each site, the benefit of each additional fish can be calculated using the power equations. In the present case, the n required has been calculated

Table 3
Required number of fish to collect per site[a] to establish a statistical difference at $\alpha = 0.05$. The calculations assumed one reference and three impacted sites.

Parameter	Amplitude of change relative to reference values	N based on reference site and predicted amplitude (data created for exposed sites) Power[b]		N using real-life data Power	
		80%	95%	80%	95%
EROD activity	3x	8	12	7	10
ECOD activity	3x	4	4	5	6
Serum SDH	2x	7	10	5	7
Naphthalene bile metabolites	4x	10	15	5	8
Pyrene bile metabolites	4x	5	6	5	6
BaP bile metabolites	4x	6	9	6	9
HSP70	2x	13	19	10	15
LSI	0.1x	43	67	13	19
GSI	0.1x	68	106	24	36
CF	0.1x	26	40	14	21

[a] These estimates used adult fish outside the reproductive activities. If a biomarker is influenced by sex, e.g. EROD activity or GSI, a larger number of individuals needs to be collected or the samples need to be segregated by sex before statistical analysis.
[b] Power: chance of obtaining a significant difference among groups at $\alpha = 0.05$.

for a power of 0.80 and 0.95, as under many situations it is prudent to reduce the possibility of Type II error where possible. From the perspective of environmental management, a Type II error is far more serious than a Type I error. A Type I error can be seen as a false alarm which could trigger further environmental protective measures – it is only a question of time before the mistake is realized through additional sampling. In contrast, a Type II error leading to a conclusion of 'no impact' would result in no remediation measures being implemented, a possible reduction in monitoring effort, and a continuing environmental deterioration. Thus, due to a lack of statistical power, there would be continued environmental degradation.

The fourth factor affecting the minimum required sample size is the variability of the parameter. Biomarkers can be notoriously variable. For example, the coefficients of variation of all parameters except CF ranged from 12.6% to 127% (Table 2), while the coefficient of variation for CF averaged 6.1%. If the variability within a sampling site is great, a larger sample size will be required to detect a given difference between means (Zar, 1996). Sources of variability for a given biomarker include individual (random) variability, systematic sampling error due to confounding factors, and analytical variability. Individual variability is often large in biological samples, especially in biochemical measurements, and it often reflects a lack of knowledge about biological or environmental factors affecting a response. Without research on those factors, the source of variation cannot be controlled, and the inherent variability might be so high that the biomarker is invalidated as part of a field monitoring program. Minimizing the effects of confounding factors can reduce systematic sampling error. For example the data set used in the present exercise included only non reproductively-active adult fish to reduce the high variability of EROD activity among female fish at the onset of spawning. Estrogen is known to down-regulate the *cyp1a* gene, so that assays of EROD activity in sexually maturing female fish approaching spawning will inflate the variance of EROD activities of a mixed sample of male and female fish (Forlin and Haux, 1990). If the biomarker selected is influenced by the gender of the fish, the data provided in Table 3 represents the number of fish *per sex* to be collected at each site, assuming that the variance is equal between sexes. It is worthwhile to note that in field studies, seasonality in biomarkers of fish health often introduces variability that is higher than inter-site variability (Hanson et al., 2010), making it increasingly difficult to relate cause and effects. A rigorous sampling program with an adequate number of fish sampled will offer a reasonable potential to offset high seasonal variability.

While the influence of confounding factors might be minimized, the analytical variability can still be surprisingly high. In an inter-laboratory round-robin, Munkittrick et al. (1993) found that EROD activities measured in sub-samples of fish livers varied considerably. For seven laboratories reporting EROD activities measured with 9000*g* supernatants (S-9 fractions), the coefficients of variation of arithmetic mean EROD activities of six fish per site sampled from reference and pulp mill sites ranged from 46–80% (calculated from Table 2, Munkittrick et al., 1993). However, the variation in induction (i.e. the proportional increase in activity between reference and exposed sites) was much less, with a cv of only 30% among the seven independent labs. This indicates that the variance among labs was likely related to differences in methods that affected induced and uninduced fish equally.

Standardization and improvement of analytical protocols can reduce analytical variability (van den Heuvel et al., 1995), thereby increasing the probability of detecting an inter-site difference. Because this variability is entirely within the control of the monitoring agency, it can be beneficial to develop Quality Assurance/Quality Control (QA/QC) protocols for each biomarker. For example, in addition to variations among fish of EROD activity, variation in EROD assays can be generated from each step of the assay, including preparation of S-9 fractions, the biochemical assay, and the analysis of data. A major source of analytical error is the difference among analysts in the efficiency of recovery of S-9 fractions from centrifuged liver homogenates. Hence, within one experiment, only one investigator should be assigned the tasks of recovering the S-9 fraction. Other sources of variation may be attributed to differences in the analyst, temperatures in the lab and in the spectrofluorometer, timing of thawing and preparation of reaction solutions, and reagent quality. Munkittrick et al. (1993) pointed out large differences among laboratories in reported extinction coefficients of standard resorufin solutions, reflecting differences among batches of standard, the instruments used to measure extinction coefficient, and the procedures of each laboratory.

To assess the occurrence and extent of variation, we maintain control sheets showing the variations among assays in activity of standard S-9 fractions, prepared from control rainbow trout or trout exposed to ß-naphthoflavone (BNF), a model CYP1A inducer. These 'lab standards' were prepared by mixing the S-9 fractions from numerous control and BNF-exposed fish, dividing the mixed S-9s into small aliquots and storing them frozen at −80 °C. One of each is analyzed with each experimental set of samples over a 6–18 month period to demonstrate that the analytical method works on each occasion, and to identify occasions when the

method generated data that might be higher or lower than normal. After a new batch of S-9s has been prepared and stored, the control chart is prepared from the first five samples of positive and negative control samples analyzed. The chart consists of the 95% confidence limits about the geometric mean EROD activity of the positive and negative controls, and of the induction (positive divided by negative). Subsequent samples are plotted on the same chart, and most of the new values should fall within the 95% confidence limits, and any random or systematic change in expected activity can be identified quickly.

As Fig. 1 demonstrates for one batch of positive and negative control S-9s tested over 16 months, that EROD activities of induced and control fish, and induction (the ratio of induced to control activities) varied considerably among assays. Because some of this variation could be due to poor mixing of the original S-9 fractions from individual fish, we also analyzed five control and five BNF S-9 standards on one occasion. The coefficient of variations for the positive and negative controls, and for induction based on arithmetic means were 31%, 19%, and 39%, respectively, much lower than the 'among assay' variations of 140%, 39%, and 104%, respectively from the first five samples tested in Fig. 1. Therefore, the scatter observed in Fig. 1 was due to 'among assay' variance rather than 'within assay' variance, and reflected differences in procedures or assay conditions, even though the analyses were done by the same person. The data also suggest an increasing trend in positive and negative control results, but not in induction. These data demonstrate the value of internal standards for detecting random or systematic changes in results of apparently 'standard' methods, and that there are no 'absolute' results for biomarker assays. Evidently, the required number of fish estimated for each biomarker (Table 3) incorporated both laboratory and inter-individual variability in the calculations.

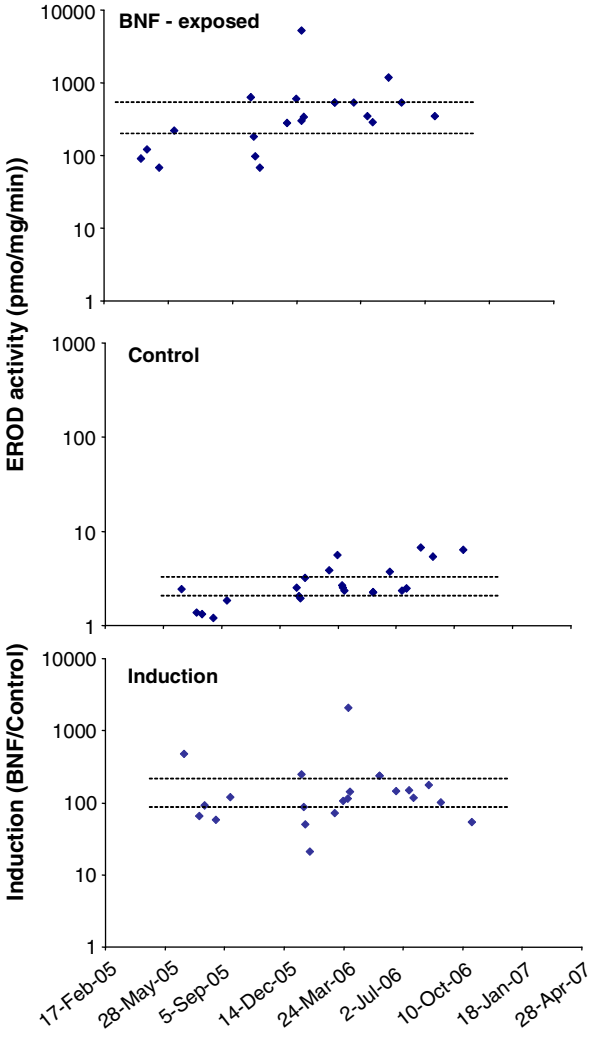

Fig. 1. EROD activity and EROD induction of individual aliquots of pooled liver S-9 fractions from ß naphthoflavone – exposed and control trout. Aliquots were thawed individually and included as QA/QC samples during EROD assays of experimental samples. Induction represents the ratio of BNF-treated to control EROD activities. The horizontal lines represent the 95% confidence limits about geometric mean values (not shown).

The large number of fish required to detect a small difference (0.1-fold change) in LSI, GSI and CF reflects the biological variability of these measurements in the fish population used for the present estimates. Some investigations have related significant biological impacts with less than 10% deviation from GSI reference conditions (Gagnon et al., 1995). However this latter study included over 3000 fish collected over three years of study (Hodson et al., 1994) which is obviously not possible for all field investigations. Fortunately, deviations in LSI and GSI from reference fish measurements are often larger than 0.1-fold (10%) in contaminated fish, making the collection of a sufficient number of fish possible for most field studies.

In the evaluation of a minimum sample size necessary to detect a statistical difference, the researcher has to decide what degree of deviation from reference conditions represents a biologically or environmentally significant difference. For a given biomarker or physiological index, the magnitude of the effects to be detected might be biologically different for individual species of fish. For example, a 2-fold increase in serum SDH activity might be related to liver damage in fish species A, while for fish species B a 5-fold increase relative to reference fish might be required before liver damage occurs.

Two important aspects have to be kept in mind when consulting the required numbers of fish suggested for any given biomarker. Firstly, the numbers presented are absolute minimum numbers of fish to obtain a statistical difference with the variability observed in a typical data set from field-collected animals. Other fish species might demonstrate higher variability and consequently, a higher number of fish will be required to demonstrate if an effect does occur. Secondly, the identification of statistical significance is in no way related to biological significance, and monitoring programs must establish on a case-by-case basis which suite of biomarkers and response sizes will be most relevant to potential cause–effect relationships.

The use of an adequate sample size for field studies can result in clearer conclusions from field investigations. It can also support permit applications for use of animals by demonstrating the minimum number of animals to be collected to achieve statistically robust outcomes. Finally, the knowledge of the minimum number of animals to be collected can in some cases contribute to environmental conservation especially when using rare and/or endangered species, as populations of fish living in severely contaminated environments are often depleted.

We found no other equivalent studies for other biomarker responses, and to our knowledge, this is the first report on minimum sample sizes for biomarkers other than EROD activity. Achieving statistical significance between sites does not help in the interpretation of the biological significance of a parameter, but well-planned field samplings will maximize the chances of correctly identifying areas of concern where remediation measures are required.

Acknowledgements

We thank Jerrold Zar, Philip Withers and Axel Buchner for their valuable guidance during the preparation of the manuscript.

References

Adams, S.M., Ham, K.D., Greely, M.S., LeHew, R.F., Hinton, D.E., Syalor, C.F., 1996. Downstream gradients in bioindicator responses: point source contaminant effects on fish health. Can J. Fish. Aquat. Sci. 53, 2177–2187.

Beliaeff, B., Burgeot, T., 1997. Sampling design optimization for EROD measurements in fish. Mar. Ecol. Prog. Ser. 153, 239–246.

Escartin, E., Porte, C., 1999. Assessment of PAH pollution in coastal areas from the NW Mediterranean through the analysis of fish bile. Mar. Pollut. Bull. 38, 1200–1206.

Fairweather, P., 1991. Statistical power and design requirements for environmental monitoring. Aust. J. Mar. Freshwater Res. 42, 555–567.

Flammarion, P., Garric, J., 1997. Cyprinids EROD activities in low contaminated rivers: a relevant statistical approach to estimate reference levels for EROD biomarker? Chemosphere 35, 2375–2388.

Flammarion, P., Garric, J., 1999. A statistical approach for classifying the extend of EROD induction of fish sampled in clean and contaminated waters. Wat. Res. 33, 2683–2689.

Forlin, L., Haux, C., 1990. Sex differences in hepatic cytochrome P-450 monooxygenase activities in rainbow trout during an annual reproductive cycle. J. Endocrinol. 124, 207–213.

Gagnon, M.M., Holdway, D.A., 2002. EROD Activity, serum SDH and PAH biliary metabolites in sand flathead (*Platycephalus bassensis*) collected in Port Phillip Bay, Victoria. Mar. Pollut. Bull. 44, 230–237.

Gagnon, M.M., Bussières, D., Dodson, J.J., Hodson, P.V., 1995. White sucker (*Catostomus commersoni*) growth, maturation and reproduction in pulp mill contaminated and reference rivers. Environ. Toxicol. Chem. 14, 317–327.

Hanson, N., Forlin, L., Larsson, A., 2010. Spatial and annual variation to define the normal range of biological endpoints: an example with biomarkers in perch. Environ. Toxicol. Chem. 29, 2616–2624.

Hodson, P.V., Dodson, J.J., Bussieres, D., Gagnon, M.M., 1994. Fish population monitoring – how many fish are enough? In: Proceedings of a Symposium on Aquatic Environmental Effects Monitoring, Quebec City, Oct. 21, 1993. Environment Canada, Ottawa, ON, Canada.

Holdway, D.A., Brennan, S.E., Ahokas, J.T., 1994. Use of hepatic MFO and blood enzyme biomarkers in sand flathead (*Platycephalus bassensis*) as indicators of pollution in Port Phillip Bay, Australia. Mar. Pollut. Bull. 28, 683–695.

Huuskonen, S., Lindström-Seppa, P., 1995. Hepathic cytochrome P4501A and other biotransformation activities in perch (*Perca fluviatilis*): the effects of unbleached pulp mill effluents. Aquat. Toxicol. 31, 27–41.

Jung, J.-H., Kim, M., Yim, U.H., Ha, S.Y., An, J.G., Won, J.H., Han, J.M., Kim, N.S., Addison, R.F.1., Shim, W.J., 2011. Biomarker responses in pelagic and benthic fish over 1 year following the Hebei Spirit oil spill (Taean, Korea). Mar. Pollut. Bull. 62, 1859–1866.

Mower, B.F., Munkittrick, K.R., McMaster, M.E., van Beneden, R.J., 2011. Response of white sucker (*Catostomus commersoni*) to pulp and paper mill effluents in the Androscoggin River, Maine, USA. Environ. Toxicol. Chem. 30, 142–153.

Munkittrick, K.R., van den Heuvel, M.R., Metner, D.A., Lockhart, W.L., Stegman, J.J., 1993. Interlaboratory comparison and optimization of hepatic microsomal ethoxyresorufin O-deethylase activity in white sucker (*Catostomus commersoni*) exposed to bleached kraft pulp mill effluent. Environ. Toxicol. Chem. 12, 1273–1282.

Oris, T.O., Roberts, A.P., 2007. Statistical analysis of cytochrome P4501A biomarker measurements in fish. Environ. Toxicol. Chem. 26, 1742–1750.

van den Heuvel, M.R., Munkittrick, K.R., Stegeman, J.J., Dixon, D.G., 1995. Second-round interlaboratory comparison of hepathic ethoxyresorufin-*o*-deethylase activity in white sucker (*Catostomus commersoni*) exposed to bleached kraft pulp mill effluent. Environ. Toxicol. Chem. 14, 1513–1520.

van der Oost, R., Beyer, J., Vermeulen, N.P.E., 2003. Fish bioaccumulation and biomarkers in environmental risk assessment: a review. Environ. Toxicol. Pharmacol. 13, 57–149.

Vijayan, M.M., Pereira, C., Kruzynski, G., Iwama, G.K., 1998. Sublethal concentrations of contaminant induce the expression of hepatic heat shock protein 70 in two salmonids. Aquat. Toxicol. 40, 101–108.

Webb, D., Gagnon, M.M., 2007. Serum sorbitol dehydrogenase activity as an indicator of chemically induced liver damage in black bream (*Acanthopagrus butcheri*). Environ. Bioindic. 2, 172–182.

Webb, D., Gagnon, M.M., Rose, T.H., 2005a. Interannual variability in fish biomarkers in a contaminated temperate urban estuary. Ecotoxicol. Environ. Saf. 62, 53–65.

Webb, D., Gagnon, M.M., Rose, T., 2005b. Metabolic enzyme activities in black bream (*Acanthopagrus butcheri*) from the Swan-Canning Estuary, Western Australia. Comp. Biochem. Physiol. C. 141, 356–365.

Zar, J.H., 1996. Biostatistical Analysis, fourth ed. Prentice Hall, Upper Saddle River, NJ, pp. 929.